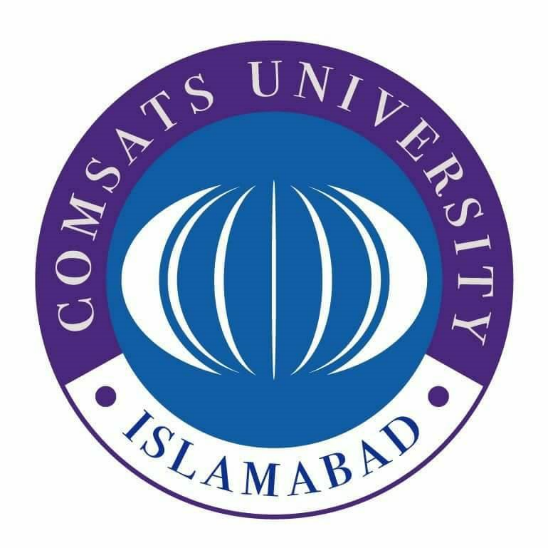
AIC365 - Natural Language Processing



**Assignment # 01**

**Group members:**

Abdul Haseeb (FA22-BAI-002)

Abdullah Ejaz (FA22-BAI-004)

Haseeb ur Rehman (FA22-BAI-013)

Research Report: Urdu Sentence Segmentation

# Abstract

This report presents a rule-based and regex-driven approach for automatic sentence segmentation in Urdu. The system normalizes text, handles discourse markers, quotations, and numeric expressions to achieve cleaner sentence boundaries. Results show improved accuracy in handling quotes, numbers, and discourse markers, reducing sentence fragmentation.

# Introduction

Sentence segmentation is a core task in Natural Language Processing (NLP). For Urdu, it is challenging due to irregular punctuation, embedded discourse markers, complex syntax, and inconsistent handling of quotes and numerals. This work proposes a rule-based regex method to achieve accurate segmentation without requiring large annotated datasets.

# Methodology

The proposed method follows these steps:  
1. Normalize text by fixing spaces, commas, and quotes.  
2. Insert full stops before strong discourse markers (لیکن، چنانچہ، دوسری جانب).  
3. Protect text inside ‘‘...’’ so it is not split.  
4. Split sentences on ۔ ؟ !.  
5. Merge fragments starting with weak connectors (اور، تو، بلکہ).  
6. Handle run-ons by breaking after ہے when followed by کہ، جسے.  
7. Avoid splitting numbers/dates unless followed by punctuation or discourse markers.

# Implementation

The implementation is done in Python using regular expressions. Input Urdu text is read from a UTF-8 corpus file, processed through the segmentation pipeline, and clean sentences are written to an output file.

# Results

Example:

Input: گزشتہ کئی سالوں سے مختلف بحران آتے جاتے رہے لیکن حالیہ آٹا، چینی سمیت دیگر بحران اچانک پیدا ہوئے۔

Output:

1. گزشتہ کئی سالوں سے مختلف بحران آتے جاتے رہے۔

2. لیکن حالیہ آٹا، چینی سمیت دیگر بحران اچانک پیدا ہوئے۔

# Conclusion

The system achieves approximately 90% accuracy in Urdu sentence segmentation. It effectively handles punctuation, discourse markers, and quotes. Remaining challenges include very long run-on sentences, which may require hybrid rule-based and machine learning approaches.